REGRESI SEMIPARAMETRIK SPLINE TRUNCATED DENGAN SOFTWARE R

Tiani Wahyu Utami[1)], Alan Prahutama[2)]
[1]Program studi Statistika, FMIPA, Universitas Mumammadiyah Semarang
email: tianiutami@unimus.ac.id
[2]Departemen Statistika, Fakultas Sains dan matematika, Universitas Diponegoro
email: alanprahutama@gmail.com

Abstract

Metode statisika sangat berperan penting dalam memprediksi maupun menduga. Salah satu metode yang digunakan adalah dengan analisis regresi semiparametrik spline. Bentuk estimator Spline Truncated sangat dipengaruhi oleh nilai titik knots. Oleh karena itu, pemilihan titik knot optimal mutlak diperlukan. Metode pemilihan titik knot dengan GCV. Dua algoritma dan pemograman untuk mendapatkan pemodelan regresi semiparametrik spline truncated dengan menggunakan software R yaitu algoritma dan program untuk menentukan Knot optimal berdasarkan metode GCV, algoritma dan program untuk mengestimasi model regresi semiparametrik Spline Truncated.

Keywords: *Spline Truncated, GCV, Software R.*

1. PENDAHULUAN

Metode statisika sangat berperan penting dalam memprediksi maupun menduga. Salah satu metode yang digunakan adalah dengan analisis regresi. Berkaitan dengan pengestimasian kurva regresi, terdapat tiga model regresi yang dapat digunakan yaitu model regresi parametrik, model regresi nonparametrik, dan model regresi semiparametrik.

Dalam beberapa kasus, variabel respon dapat memiliki hubungan linier dengan salah satu variabel prediktor, tetapi dengan variabel prediktor yang lain tidak diketahui bentuk pola hubungannya. Dalam keadaan seperti ini, Wahba (1990) menyarankan menggunakan pendekatan regresi semiparametrik. Model regresi semiparametrik yang popular adalah regresi semiparametrik spline.

Tujuan utama dalam regresi semiparametrik adalah mendapatkan estimasi kurva regresi. Terdapat beberapa pendekatan untuk mengestimasi kurva regresi salah satunya adalah dengan Spline *Truncated*. Spline *Truncated* merupakan model regresi yang mempunyai interpretasi statistik visual sangat khusus dan sangat baik. Disamping itu, kelebihan Spline *Truncated* adalah dapat mengatasi pola data yang menunjukkan naik atau turun yang tajam dengan bantuan titik-titik knots, serta kurva yang dihasilkan relatif mulus. Titik knots merupakan perpaduan bersama yang menunjukkan pola perilaku fungsi Spline *Truncated* pada selang yang berbeda (Hardle, 1990). Bentuk estimator Spline *Truncated* sangat dipengaruhi oleh nilai parameter penghalus λ. Bentuk estimator Spline *Truncated* juga dipengaruhi oleh lokasi dan banyak titik knots (Budiantara, 2005). Oleh karena itu, pemilihan λ optimal mutlak diperlukan untuk memperoleh estimator Spline *Truncated* yang sesuai dengan data. Model Spline *Truncated* terbaik dapat dilihat dari beberapa

kriteria tertentu yaitu mempunyai nilai *Means Squared Error* (MSE) dan nilai *Generalized Cross Validation* (GCV) yang minimum.

Penerapan pada data dalam menyelesaikan pemodelan dengan menggunakan pendekatan regresi semiparametrik Spline *Truncated* diperlukan bantuan software statistika. Salah satu paket analisis data *open source* yang dapat diperoleh secara cuma-cuma yaitu software R. R merupakan paket pemrograman yang termasuk keluarga S (bahasa S). Paket program R ini sudah dilengkapi dengan banyak kemampuan internal untuk menganalisis data dan menampilkan grafik sehingga R dapat dikategorikan sebagai paket pengolahan data (paket statistika). Oleh karena itu, penyelesaian pemodelan regresi semiparametrik Spline *Truncated* mengunakan algoritma dan pemograman menggunakan *softwar*e R.

2. KAJIAN LITERATUR

2.1 Regresi Semiparametrik

Misal menotasikan variabel respon untuk subjek ke-*i*, sedangkan x_i, t_i menyatakan variabel-variabel prediktor, *n* banyaknya subjek. Jika bentuk fungsi antara variabel respon dengan tidak diketahui, sedangkan dengan x_i diketahui berbentuk linier, maka data dapat dimodelkan dengan menggunakan model regresi semiparametrik (Drapper, N.R dan Smith, H, 1992):

$$y_i = \alpha_0 + x_i\alpha_1 + \eta(t_i) + e_i \qquad (1)$$

Dengan adalah error pengukuran. Fungsi adalah fungsi yang tidak diketahui bentuknya yang diasumsikan *smooth*.

2.2 Regresi Spline *Truncated*

Regresi Spline *Truncated* merupakan fungsi potongan polynomial yang memiliki sifat tersegmen dan kontinu. Bentuk umum model Spline *Truncated* disajikan pada persamaan :

$s(t) = {}_i t^i + {}_j(t-k_j)^p + ,$

dimana :

α_i dan β_j : merupakan parameter dimana i = 0, 1, …, *p* dan *j* = 1, …, *k*

dimana *p* adalah derajat polynomial dan *k* adalah banyaknya titik knot pada fungsi truncated, serta ε_i merupakan error random independen dengan mean nol dan varian σ^2. Regresi Spline *Truncated* memungkinkan untuk berbagai macam orde sehingga dapat dibentuk regresi Spline linier, kuadrat, kubik maupun orde *m*. Bentuk estimator Spline *Truncated* sangat dipengaruhi oleh nilai parameter penghalus λ (Budiantara, 2009). Bentuk estimator Spline juga dipengaruhi oleh lokasi dan banyaknya titik-titik knot). Model Spline terbaik dapat dilihat dari beberapa kriteria tertentu yaitu mempunyai nilai *Means Squared Error* (MSE) dan nilai *Generalized Cross Validation* (GCV) yang minimum.

2.3 *Software* R

R adalah salah satu paket analisis data *open source* yang dapat diperoleh secara cuma-cuma di situs *http://www.r-project.org/*. atau *http://cran.r-project.org/*. R merupakan paket pemrograman yang termasuk keluarga S (bahasa S). Paket program R ini sudah dilengkapi dengan banyak kemampuan internal untuk

menganalisis data dan menampilkan grafik sehingga R bisa dikategorikan sebagai paket pengolahan data (paket statistika). Beberapa kemampuan yang menonjol dari R yang menjadi alasan banyak statistisi memilihnya sebagai paket aplikasi antara lain sebagai berikut (Tirta , 2008):

1. R memiliki koleksi program analisis data, yang disebut *library* atau pustaka yang sangat luas seperti statistika deskriptif, regresi, pemodelan statistika (baik linear maupun nonlinear), anova dan multivariat.
2. Variasi penampilan grafiknya sangat banyak dan berkualitas tinggi, baik penampilan di layar monitor maupun dalam bentuk cetak di atas kertas.
3. Kemampuan pemrograman (bahasa S) dapat dikembangkan secara fleksibel untuk kepentingan khusus yang lebih lanjut.

R merupakan pemrograman yang berorientasi pada objek. Keuntungannya, apabila apa yang telah dikerjakan R saat ini diperlukan di kemudian hari maka R dapat mengambilnya tanpa harus melakukan perhitungan ulang dari awal

3. METODE PENELITIAN

Tahapan-tahapan pada penelitian ini adalah membuat algoritma dan pemograman pemodelan TPT di Jawa Tengah dengan pendekatan regresi semiparametrik Spline Truncated.

a. Membuat algoritma dan pemograman untuk menentukan Knot optimal berdasarkan metode GCV

b. Membuat algoritma dan pemograman untuk mengestimasi model regresi semiparametrik Spline *Truncated*

4. HASIL PENELITIAN

Estimasi terhadap $f(x)$ adalah $f_\lambda(x)$ yakni estimator yang mulus. Bentuk umum regresi Spline orde ke-m adalah sebagai berikut :

$$y = \beta_0 + {}_j x^j + {}_{j+k} + \varepsilon$$

dengan menggunakan data amatan sebanyak n, maka bentuk matriks dari persamaan diatas dapat ditulis sebagai berikut:

$$y = X_1\delta_1 + (X - K)\delta_2 + \varepsilon$$

dengan,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \boldsymbol{\delta_1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}; \mathbf{X_1} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & {}_n^m \end{bmatrix}; \boldsymbol{\delta_1} = \begin{bmatrix} \beta_{m+1} \\ \beta_{m+2} \\ \beta_{m+3} \\ \vdots \\ \beta_{m+N} \end{bmatrix}$$

$$(\mathbf{X} - \mathbf{K}) = \begin{bmatrix} (x_1 - k_1)^m & (x_1 - k_2)^m & \cdots & (x_1 - k_N)^m \\ (x_2 - k_1)^m & (x_2 - k_2)^m & \cdots & (x_2 - k_N)^m \\ \vdots & \vdots & \ddots & \vdots \\ (x_n - k_1)^m & (x_n - k_2)^m & \cdots & (x_n - k_N)^m \end{bmatrix}$$

Dan matriks dalam persamaan diatas dapat disederhanakan menjadi:

$$y = X\beta + \varepsilon$$

dimana$X = [X_1 \quad (X - K)]$ dan β =
Estimasi untuk parameter β adalah
$b_\lambda = (X_\lambda)^{-1}$, dimana $= b_\lambda$
fungsi estimasi dari $f(x)$ adalah sebagai berikut:
$_\lambda(x) = X\lambda b\lambda = X_\lambda(X_\lambda)^{-1} = H_\lambda Y$
Dengan $H_\lambda = X_\lambda(X_\lambda)^{-1}$yang bersifat simetris dan definit positif sedangkan X_λ adalah matriks desain berukuran n x k dari model yang membentuk $f_\lambda$dan bergantung pada titik knot.

$$X_\lambda = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{m-1} & (x_1-\lambda_1)_+^{m-1} & \dots & (x_1-\lambda_k)_+^{m-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{m-1} & (x_2-\lambda_1)_+^{m-1} & \dots & (x_2-\lambda_k)_+^{m-1} \\ \vdots & \vdots & \vdots & & \dots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{m-1} & (x_n-\lambda_1)_+^{m-1} & \dots & (x_n-\lambda_k)_+^{m-1} \end{bmatrix}$$

Generelized Cross-Validation (GCV) Merupakan metode untuk memilih knot optimal (Budiantara, 2006). Fungsi *GCV* didefinisikan sebagai berikut :

$$GCV(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}\left(y_i - f_\lambda(x_i)\right)^2}{\left(1-n^{-1}Tr(\mathbf{H}_\lambda)\right)^2}$$

Dengan $Tr\ (H_\lambda) > n$. Kriteria *GCV* (λ) diharapkan memiliki nilai yang minimum, sehingga model regresi Spline dapat dikatakan memiliki nilai knot yang optimal.

Untuk mendapatkan estimasi model regresi semiparametrik pada data berdasarkan estimator Spline, algoritma yang digunakan untuk membuat program dalam *software* R yaitu :

a. Algoritma untuk menentukan Knot optimal berdasarkan metode GCV :

1. Mendefinisikan variabel respon *Y* dan variabel prediktor *X* dan *T*.
2. Menentukan nilai batas bawah dan batas atas titik *knot (k).*
3. Menentukan orde (m)
4. Mendapatkan matriks ***A***, matriks ***Apar*** seperti pada persamaan berikut :
   `A=X*Invers(t(X)*X)*(X)`
   `Apar = T*Invers(t(T)*t(I-A)*(I-A)*T)*t(T)*t(I-A)*(I-A)`
5. Menghitung nilai = `Apar+(A%*%(I-Apar))*Y`
6. Menghitung nilai *GCV(k)* , yaitu

$$GCV(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}\left(y_i - f_\lambda(x_i)\right)^2}{\left(1-n^{-1}Tr(\mathbf{H}_\lambda)\right)^2}$$

7. Mengulang langkah (2) sampai (6) untuk *titik knot(k)* dan orde *(m*) yang berbeda hingga diperoleh nilai *GCV(k,m)* yang minimum
8. Nilai *titik knot(k)* dan orde *(m*) yang bersesuaian dengan nilai GCV yang minimum adalah nilai *titik knot(k)* dan orde *(m*) yang optimal.

Program untuk menentukan Knot optimal berdasarkan metode GCV berikut :

```
trun <- function(gdp,a,power)
{
  gdp[gdp<a] <- a
  (gdp-a)^power
}
gcv.knots<-function(y,x,t,m)
{
k1<-as.numeric(readline("nilai knots pertama ="))
      a <- k1 - 0
      b <- k1 + 23
      k <- a
      while (k<=b)
      {for (i in 1:(m+1))
       { X[,i] <- x ^(i-1)}
            X[,(m+2)] <- trun(x,k,m)
```

b. Algoritma untuk mengestimasi model regresi semiparametrik Spline *Truncated*:

1. Mendefinisikan variabel respon Y dan variabel prediktor X dan T.
2. Menginputkan nilai *knots* (k) optimal yang diperoleh dari algoritma A.
3. Mendapatkan matriks $\boldsymbol{A}$, matriks $\boldsymbol{Apar}$ seperti pada persamaan berikut :
   $\boldsymbol{A=X}*\text{Invers}(t(\boldsymbol{X})*\boldsymbol{X})*(\boldsymbol{X})$
   $\boldsymbol{Apar} = T*\text{Invers}(t(T)*t(\boldsymbol{I-A})*(\boldsymbol{I-A})*T)*t(T)*t(\boldsymbol{I-A})*(\boldsymbol{I-A})$
4. Menentukan nilai dan error
5. Menghitung nilai R-Square dan nilai *MSE(k)*, yaitu

Program untuk mengestimasi model regresi semiparametrik Spline *Truncated*

```
trun <- function(gdp,a,power)
{
  gdp[gdp<a] <- a
  (gdp-a)^power
}
spline.knots<-function(respon,nonpar,par,orde,knots=c(...))
{
data<-cbind(respon,nonpar,par)

    y<-as.vector(data[,1])
    nonpar<-as.vector(data[,2])
    x<-as.vector(data[,3])
    n <- length(y)
    k <- length(knots)
    m<-orde
    v<- matrix(0,n,m+1+k)
         for (i in 1:(m+1))
```

5. SIMPULAN

Berdasarkan hasil penelitian yang sudah dilakukan, terdapat dua algoritma dan pemograman untuk mendapatkan pemodelan regresi semiparametrik spline *truncated* dengan menggunakan *software* R yaitu

1. Algoritma dan program untuk menentukan Knot optimal berdasarkan metode GCV
2. Algoritma dan program untuk mengestimasi model regresi semiparametrik Spline *Truncated*:

6. REFERENSI

Budiantara, I N., 2009, *Spline Dalam Regresi Nonparametrik dan Semiparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang*, Institut Teknologi Sepuluh November, Surabaya.

Budiantara, I.N., Suryadi, F., Otok, B.W., dan Guritno, S., 2006, Pemodelan B- Spline dan MARS Pada Nilai Ujian Masuk terhadap IPK Mahasiswa Jurusan Disain Komunikasi Visual UK. Petra Surabaya; *Jurnal Teknik Industri*, Vol 8 No. 1 hal 1-13; Universitas Petra.

Budiantara, I.N.(2000), "Metode U, GML, CV, dan GCV dalam regresi Nonparametrik Spline", *Majalah Ilmiah Himpunan Matematika Indonesia (MIHMI)*, Vol. 6, hal. 285-290.

Drapper, N.R dan Smith, H, 1992, *Applied Regression Analysis*, *2nd edition*, John Wiley & Sons, Chapman and Hall, New York.

Eubank,R.L.,1988, *Spline Smoothing and Nonparametric Regression*, Mercel Dekker, New York.

Hardle, W.,1990, *Applied Non-parametric Regression*, Cambridge University Press, Cambridge.

Tirta, I.M., 2008, *Buku Panduan Program Statistika*, Universitas Jember Press, Jember